PHIL 7001: Homework Assignment 2

Professor Boris Babic

Instructions: This is the second homework assignment for PHIL 7001. The assignment will be graded out of a total of 100 points. It is open notes, and open book – this means that you can use any resources you like in order to formulate your answer. You are also allowed to use any calculator you wish. However, you do have to work on this assignment on your own/individually.

You will have seven days to complete this assignment. Please be careful when completing your online assignment to select the intended answer or input the intended value. We will not be able to regrade accidental or mistaken answers.

Logistic Regression and Classification (40 points)

1. For the following scenarios, determine whether logistic or linear regression is more suitable (20 points, 4 points each).

   (a) You are tasked with predicting the grade that students will receive on a final exam on the basis of the number of hours they studied for that exam.

   (b) You wish to build a model to classify emails as either spam or not spam on the basis of various email characteristics.

   (c) You work in a hospital and want to develop a model which will predict the probability that a patient has a specific medical disease on the basis of a set of lab test results.

   (d) A research project aims to predict the expected increase in blood pressure on the basis of a person's age and the amount of fat calories they consume on a daily basis.

   (e) A retail store wants to estimate the likelihood that a particular product will go out of stock within the next month based on historical sales data and product characteristics.

2. Which of the following best describes the decision boundary in logistic regression? (4 points)

(a) It is the threshold for classification.

(b) It is a linear line that separates data points.

(c) It is the sum of weighted features.

(d) It is the ROC curve.

3. **True or False**: Logistic regression is a type of supervised learning (4 points)

4. In a confusion matrix, which metric represents the number of positive cases correctly identified by the model? (4 points)

   (a) True Positives (TP)

   (b) False Positives (FP)

   (c) False Negatives (FN)

   (d) True Negatives (TN)

5. In a confusion matrix, which metric represents the number of negative cases incorrectly identified by the model? (4 points)

   (a) True Positives (TP)

   (b) False Positives (FP)

   (c) False Negatives (FN)

   (d) True Negatives (TN)

6. True or False: An AUC of 0.5 in the ROC curve indicates a perfect classifier. (4 points)

Decision Trees and Random Forests (20 points, 4 points each)

1. Random Forests are considered a ________ (supervised/unsupervised) machine learning model.

2. Random Forests are considered an ensemble machine learning model? (true or false)

3. What is the primary advantage of using a random forest over a single decision tree?

   (a) Random forests are always faster to train.

   (b) Random forests can handle only linear data.

   (c) Random forests can reduce overfitting and improve accuracy.

   (d) Random forests are interpretable.

4. In a random forest, how are individual decision trees generated?
   (a) Each tree is trained on the entire dataset.
   (b) Each tree is trained on a subset of the data.
   (c) All trees are identical to each other.
   (d) Trees are trained sequentially.
5. In the random forest package in R, you are required to choose between a radial or linear kernel as one of the tuning parameters. True or false.

SVM (20 points, 4 points each)

1. SVM is capable of creating non linear decision boundaries (true or false)
2. SVM is considered a supervised/unsupervised machine learning model (fill in the blank).
3. SVM is considered an ensemble machine learning model? True or false.
4. What is the primary goal of an SVM when finding a decision boundary?
   (a) Maximize the margin between classes.
   (b) Minimize the number of support vectors.
   (c) Create a non-linear decision boundary.
   (d) Overfit the training data.
5. What is the ”margin” in an SVM?
   (a) The distance between the decision boundary and the nearest data point
   (b) The width of the decision boundary
   (c) The number of support vectors
   (d) The number of features in the dataset

Evaluating Model Performance (20 points, 5 points each)

1. You are given the following information from a medical test. Please calculate the sensitivity score of this test. Please round your answer to three decimal places.
   - True Positives (TP): 75

- False Positives (FP): 15
- True Negatives (TN): 120
- False Negatives (FN): 10

2. In a classification problem, you have observed the following performance from a certain model. Calculate this model's specificity score. Please round your answer to three decimal places.

   - True Positives (TP): 90
   - False Positives (FP): 20
   - True Negatives (TN): 65
   - False Negatives (FN): 5

3. You are given the following information from a medical test. Calculate the precision associated with this test. Please round your answer to three decimal places.

   - True Positives (TP): 120
   - False Positives (FP): 30
   - True Negatives (TN): 160
   - False Negatives (FN): 10

4. You are given the following information from a medical test. Calculate the recall score associated with this test. Please round your answer to three decimal places.

   - True Positives (TP): 50
   - False Positives (FP): 10
   - True Negatives (TN): 30
   - False Negatives (FN): 5